AMD

Accelerate Your Discoveries in HPC & AI with AMD

Advanced Modeling & Simulation (AMS) Seminar Series

NASA Ames Research Center, June 29, 2021

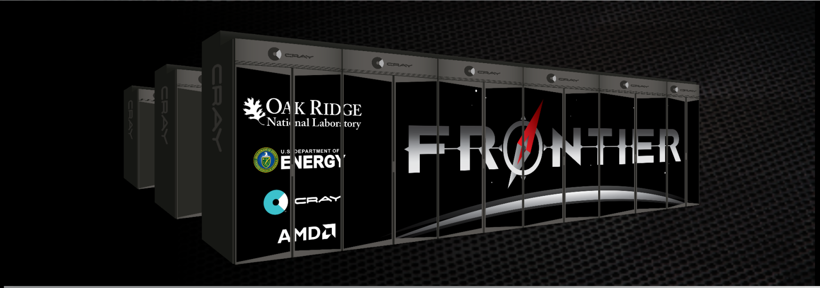
CAUTIONARY STATEMENT

This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, timing and expected benefits of AMD products and product roadmaps, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Investors are urged to review in detail the risks and uncertainties in AMD’s Securities and Exchange Commission filings, including but not limited to AMD’s most recent reports on Forms 10-K and 10-Q.

AMD does not assume, and hereby disclaims, any obligation to update forward-looking statements made in this presentation, except as may be required by law.

AMD EPYC INSTINCT

LEADING THE NEXT-GEN SUPERCOMPUTING & EXASCALE ERA



- Powered by AMD EPYC™ CPUs & AMD Instinct™ GPUs
- ~1.5 ExaFLOPS expected
- Expected to be more powerful than today's top 50 Fastest supercomputers combined
- Shipment in 2021

- Powered by next gen AMD EPYC™ CPUs & AMD Instinct™ GPUs
- ~2.0 ExaFLOPS expected
- Expected to be more powerful than today's 200 fastest supercomputers combined
- Shipment in 2023

VETENSKAP OCH KONST KUNGL TEKNISKA HÖGSKOLAN

LUMI

CGG GeoConsulting

EuroHPC Joint Undertaking

Nikhef

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

PAWSEY supercomputing centre

Source: https://www.top500.org/lists/2020/11/

Use of third-party marks/logos/products is for informational purposes only and no endorsement of or by AMD is intended or implied. GD-83

APPLICATION OPTIMIZED ARCHITECTURES

HIGHEST EFFICIENCY THROUGH DOMAIN SPECIFIC OPTIMIZATION

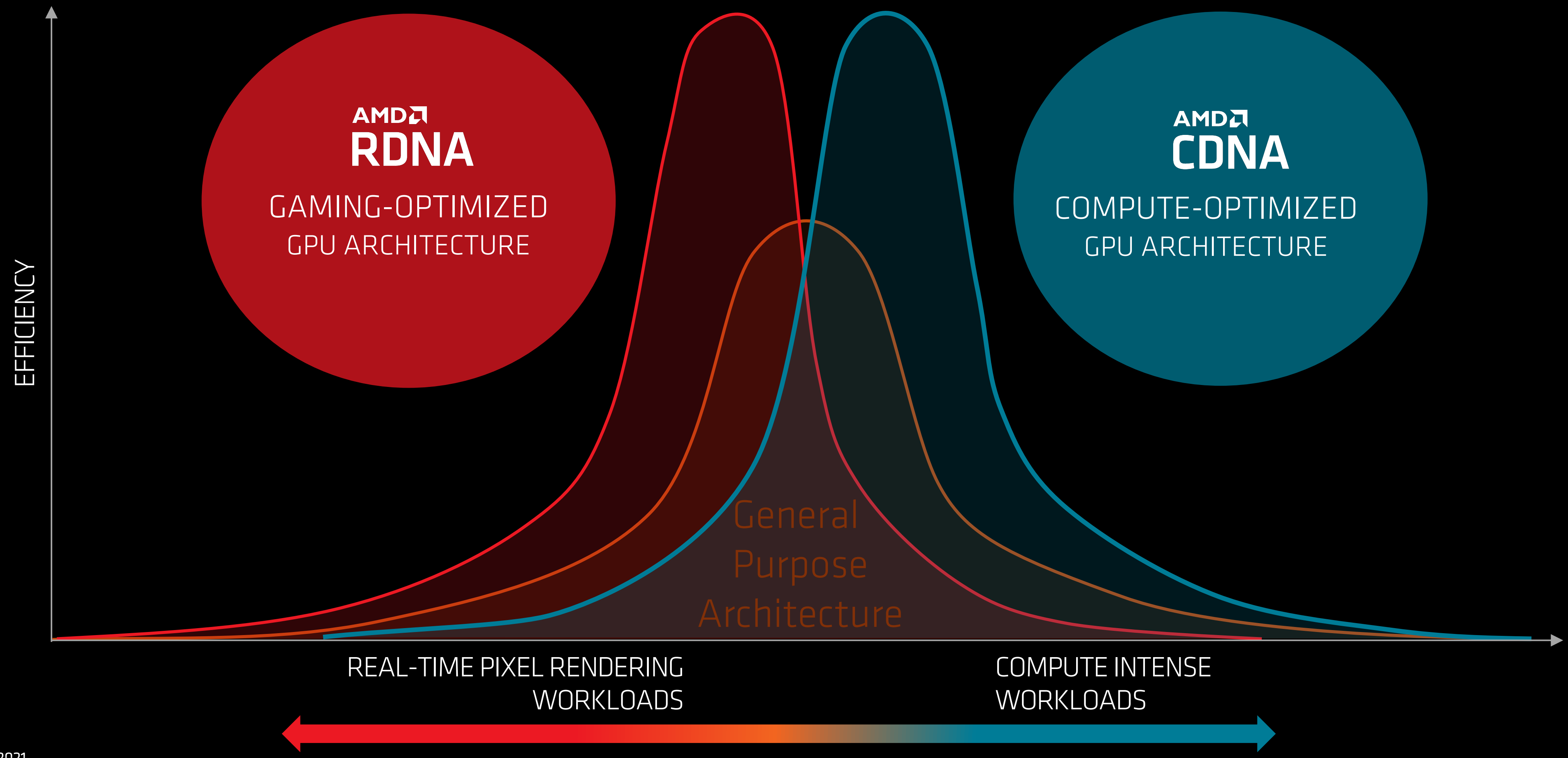


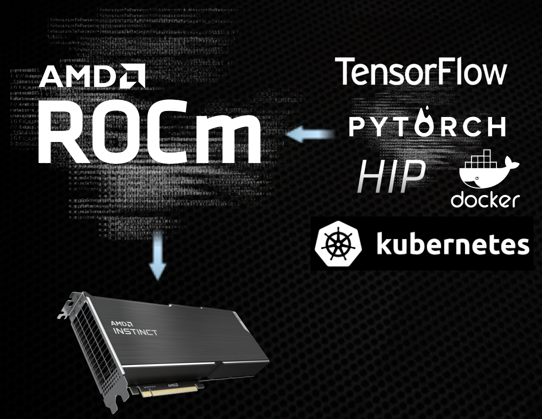
CHART FOR ILLUSTRATIVE PURPOSES

AMD PLATFORM FOR ACCELERATED COMPUTING

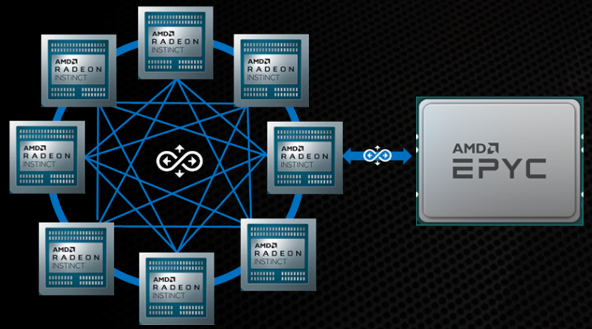
SETTING THE BAR FOR THE FUTURE OF HPC & AI

AMD CDNA

DOMAIN-OPTIMIZED ARCHITECTURE



OPEN & PORTABLE SOFTWARE



UNIFIED CPU & GPU PLATFORM

AMD

DATA CENTER GPU ARCHITECTURE ROADMAP

7nm

GCN

First 7nm
Data Center GPU

AMD Radeon Instinct™ **MI50**

7nm

AMD **CDNA**

2nd Gen AMD Infinity Architecture
Optimized for ML/HPC

AMD Instinct™ **MI100**

Advanced Node

AMD **CDNA 2**

3rd Gen AMD Infinity Architecture
Extends to Exascale

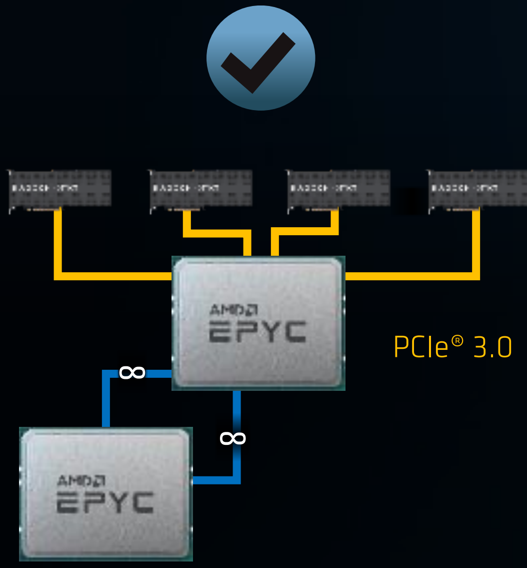
"*MINEXT*" coming by year end 2021

2019 — 2022

Roadmaps Subject to Change

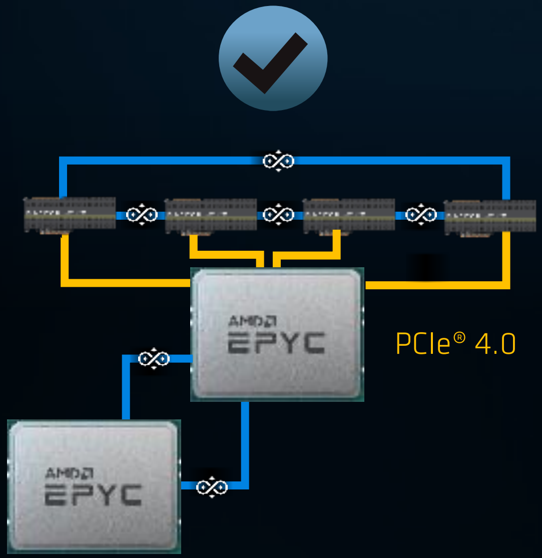
AMD INFINITY ARCHITECTURE ROADMAP

INNOVATION THROUGH CLOSER CPU AND GPU INTEGRATION



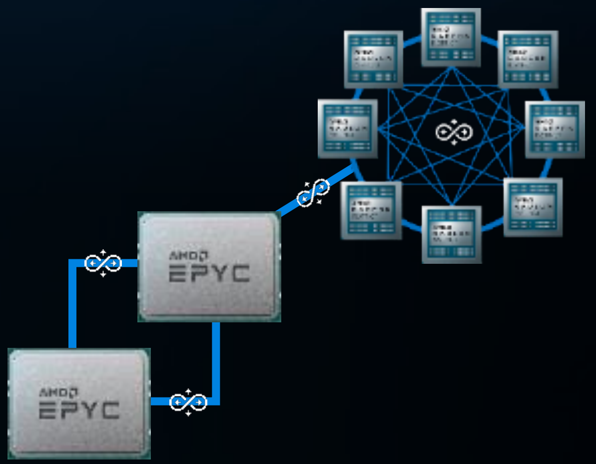
CPU CONNECTIVITY

1st Gen
AMD Infinity Fabric™



4-WAY GPU CONNECTIVITY

2nd Gen
AMD Infinity Architecture



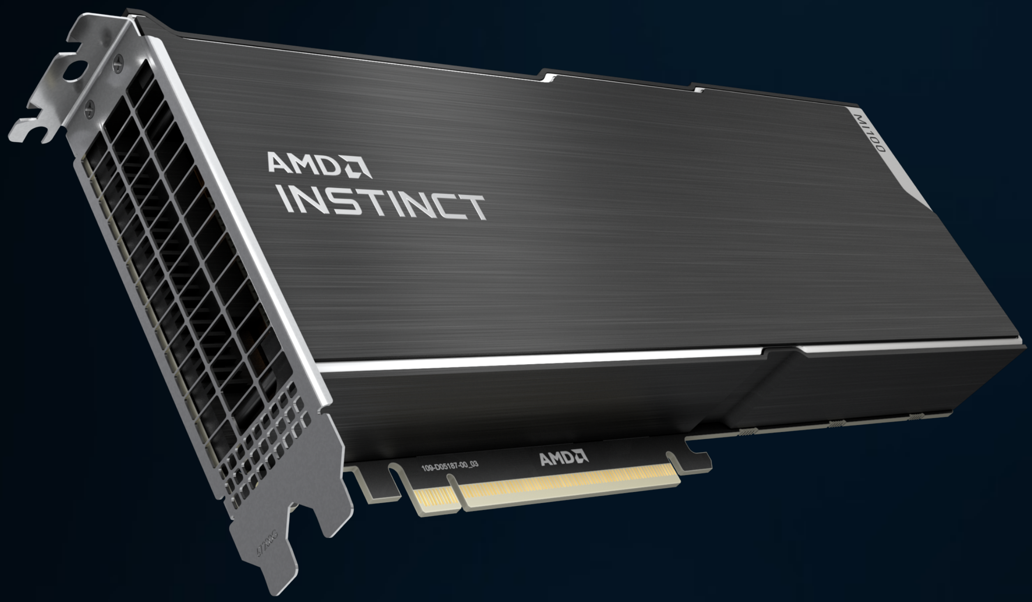
UP TO 8-WAY GPU WITH COHERENT CONNECTIVITY

3rd Gen
AMD Infinity Architecture

2017 — *2022*

Roadmap Subject to Change

AMD



AMD INSTINCT™ MI100 GPU

THE WORLD'S FASTEST HPC GPU

UP TO
11.5TF First AMD Data Center GPU to Surpass 10TF FP64 Barrier

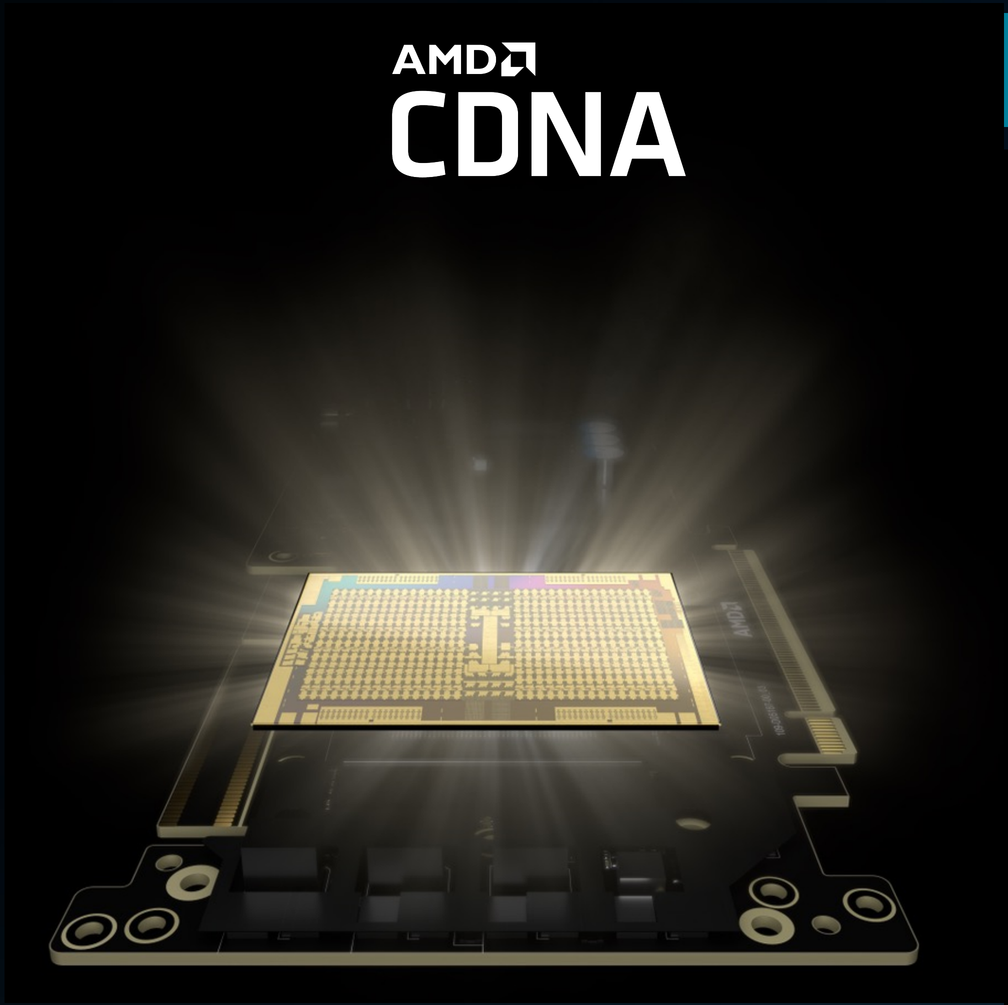
NEARLY
3.5X Faster FP32 Matrix Math than Prior Gen MI50

NEARLY
7X On AI Workloads with Mixed Precision and FP16 than Prior Gen MI50

SEE ENDNOTES MI100-04, MI100-17

AMD CDNA™ ARCHITECTURE

The All-New AMD GPU Architecture for the Modern Era of HPC & AI



ARCHITECTED WITH A FOCUS ON COMPUTE

- **AMD CDNA Architecture Overview**
  Enhanced Compute Units with Matrix Core Engine to boost computational throughput for FP32, FP16 numerical functions
- **L2 Cache and Memory**
  The 8MB L2 cache is shared across the whole chip and physically partitioned into 32 slices with a total aggregate bandwidth of 3TB/s

  The 32GB of HBM2 comes in four 8-high stacks for an aggregate theoretical throughput of 1.23TB/s
- **Communication and Scaling**
  The AMD Infinity Fabric™ links operate at 23GT/s and are 16-bits wide offering full connectivity in quad GPU configurations

  The Infinity Fabric links support coherent GPU memory, which enables multiple GPUs to share an address space and tightly cooperate on a single problem.

Source: AMD CNDA Architecture Whitepaper

AMD

AMD INSTINCT™

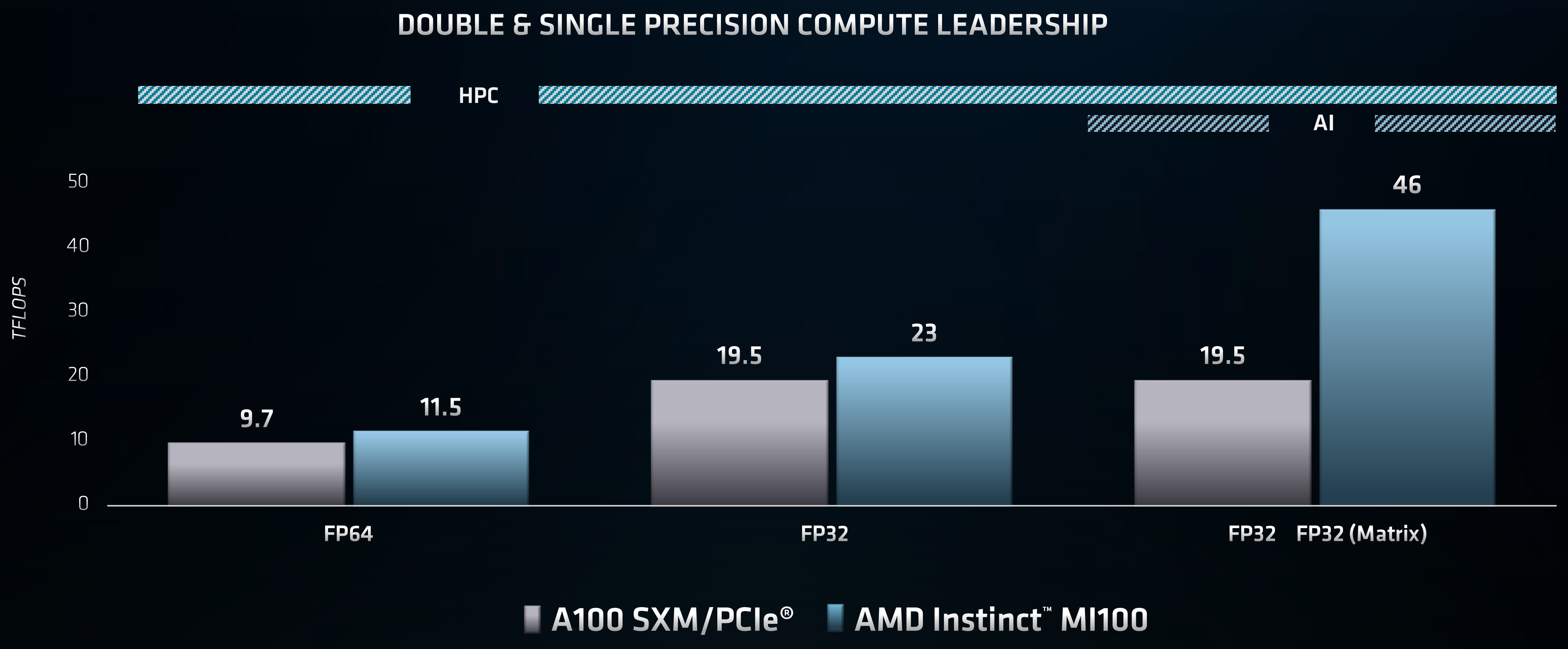
MI100

COMPUTE UNITS	120
FMA64 & FP64 PEAK	*Up to* **11.5** TFlops
FMA32 & FP32 PEAK	*Up to* **23.1** TFLops
FP32 MATRIX PEAK	*Up to* **46.1** TFLops
FP16 MATRIX PEAK	*Up to* **184.6** TFLops
BFLOAT16 PEAK	*Up to* **92.3** TFLops
MEMORY SIZE	**32GB** HBM2
MEMORY BANDWIDTH	**1.23** TB/s
PCIE® SUPPORT	**Gen4**
INFINITY FABRIC™ LINKS (X16)	*Up to* **276 GB/Sec** (Peak I/O Bandwidth)
MAX POWER / CONNECTORS	Up to **300W**

SEE ENDNOTES CDNA-04, MI100-03, MI100-07, MI100-08.

AMD INSTINCT™ MI100: THE NEW HPC GPU LEADER

SETTING NEW BAR FOR PERFORMANCE



SEE ENDNOTES MI100-03

AMD

AMD INFINITY FABRIC™ TECHNOLOGY
HIGH SPEED P2P INTERCONNECT

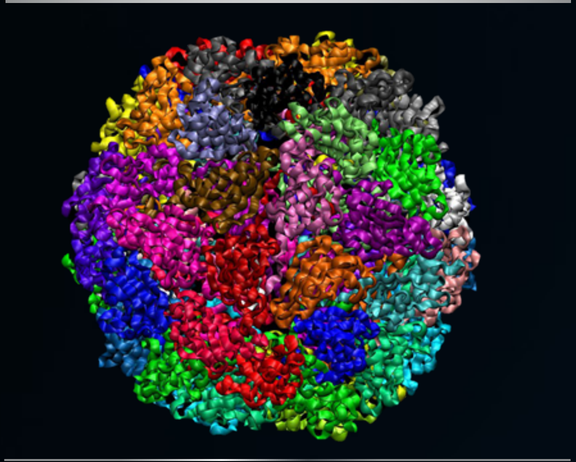
~2x higher throughput vs PCIe® 4.0 w/ 3x Infinity Fabric links per GPU

Fully-connected quad GPU hives with up to 552 GB/s peer-to-peer peak I/O bandwidth for HPC & ML

Two groups ("Hives") of 4x GPUs connected via PCIe

AMD INSTINCT™ MI100 GPU

POWERING EARLY EXASCALE SCIENCE AT OAK RIDGE



NAMD

Molecular Dynamics

~3x Faster vs V100

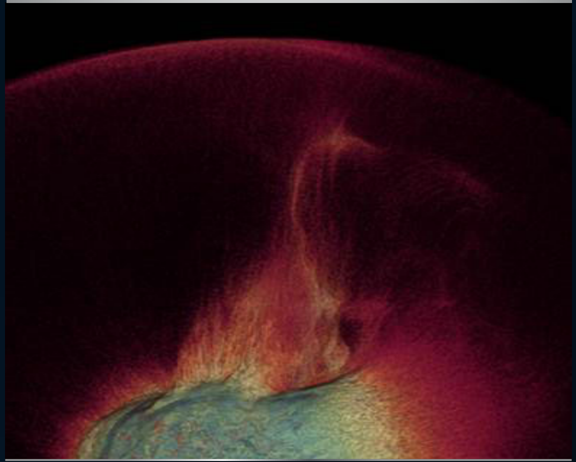
Dr. Emad Tajkhorshid, University of Illinois at Urbana-Champaign
Josh Vermaas, Computational Scientist
Arnold Tharrington, Computational Scientist

CHOLLA

Formation of Galaxies

1.4x Faster vs V100

Evan Schneider, Assistant Professor, Physics & Astronomy
Reuben Budiardja, Computational Scientist



PIConGPU

Laser Radiation Cancer Therapies

1.4x Faster vs MI60

René Widera, Computational Scientist
Sunita Chandrasekaran, Assistant Professor of Computer Science

GESTS

Fluid Turbulence

2.6x Faster vs V100

Stephen Nichols, Computational Scientist
P.K. Yeung, Professor of Aerospace Eng

SOURCE: OAK RIDGE NATIONAL LABORATORY: NAMD 2.14, STMV 1.06M ATOMS BENCHMARK, 2X EPYC 7742 + MI100 VS 2X POWER9 + V100 SXM, CHOLLA, TOTAL RUN MEASURED. 2X EPYC 7742 + MI100 VS 2X EPYC 7742 + V100, PICONGPU, TOTAL RUN MEASURED. 2X EPYC 7742 + MI100 VS 2X EPYC 7742 + V100, GESTS, TOTAL RUN MEASURED, 2X EPYC 7742 + MI100 VS 2X EPYC 7742 + V100

AMD ROCm

THE FIRST

OPEN SOFTWARE PLATFORM FOR GPU COMPUTE

Unlocked GPU Power To Accelerate Computational Tasks	Optimized for HPC and Deep Learning at Scale	Enabling Innovation, Collaboration, and Efficiency

Complete Solutions for HPC/AI Workloads

AMD EPYC | INSTINCT + AMD ROCm

Planned for June 2021

AMD Infinity Hub
Containerized HPC Apps and ML Frameworks
docker

Single place for <u>researchers, data scientists and end-users</u> to easily find, download and install containerized HPC apps and ML frameworks that are optimized and supported on ROCm™

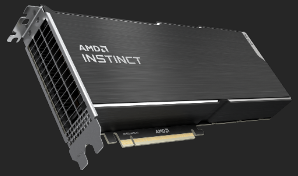
AMD ROCm
Compilers, Dev Tools, Libraries, Mgmt Tools

Open software platform for <u>developers</u> to port or build high-performance applications that run on any CPU/GPU

Validated, Optimized Systems
Hewlett Packard Enterprise
DELL EMC
SUPERMICRO
GIGABYTE™

Range of system types & configurations from leading OEMs/ODMs taking advantage of AMD Infinity Architecture

AMD CDNA



AMD INSTINCT MI 100

Purpose-built accelerators for HPC/AI workloads

AMD ROCm™ 4.0:

PRE-EXASCALE STACK FOR HPC & ML

AMD ROCm

Open Source & Portable

Unlocks full system performance

Datacenter-Ready at Scale

Applications	High Performance Computing		Machine Learning	
Cluster Deployment	Singularity	SLURM	Docker	Kubernetes
Tools	Debugger	Profiler, Tracer	System Valid.	System Mgmt.
Portability Frameworks	Kokkos	RAJA	TensorFlow	PyTorch
Math Libraries	RNG, FFT	Sparse	BLAS, Eigen	MIOpen
Scale-out Comm. Libraries	OpenMPI	UCX	MPICH	RCCL
Programming Models	OpenMP	HIP	OpenCL™	Python
Drivers	RedHat, CentOS, SLES & Ubuntu Device Drivers and Run-Time			
Processors	CPU + GPU			

Beta/Early | Production

CUSTOMER CASE STUDIES

z = 0.42 DM Density Gas Density

ORNL Advance Galaxy Simulations with AMD Instinct™ MI100 GPUs

Accelerating galaxy formation research at unprecedented scale and resolution using AMD powered supercomputing

AMD EPYC | INSTINCT

CUSTOMER

Oak Ridge National Laboratory

INDUSTRY

Scientific research

CHALLENGES

Increasing scale and resolution with numerical simulation of galactic evolution with CHOLLA software

SOLUTION

Deploy AMD Instinct MI100 GPUs with ROCm open software to prepare for Frontier supercomputer

RESULTS

Easy porting to AMD Instinct GPUs with

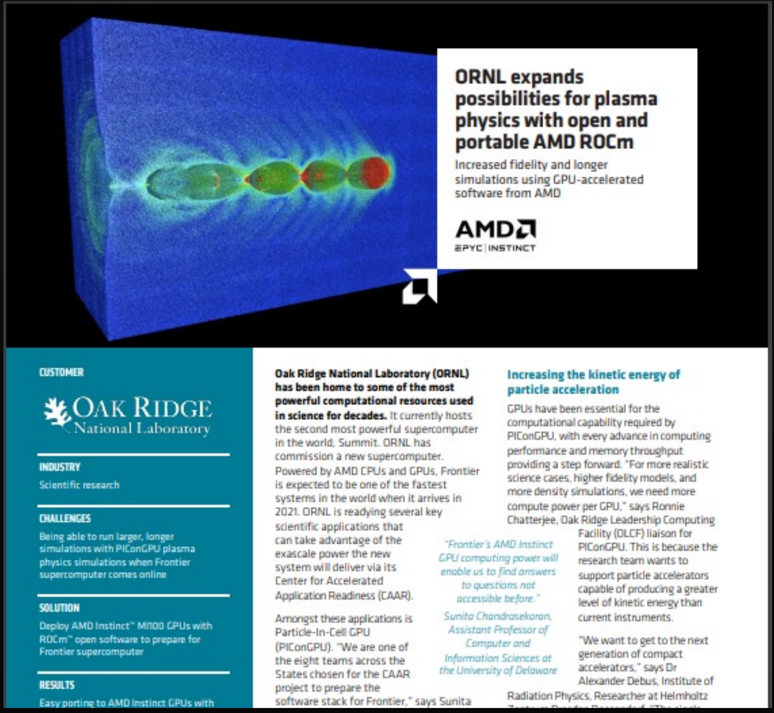
Supercomputing is revolutionizing science, with the fastest systems in the world providing unparalleled path to discovery. Oak Ridge National Laboratory (ORNL) is at the forefront of this trend, and already hosts one of the fastest supercomputer in the world. But ORNL has commissioned a new supercomputer for its Leadership Computing Facility (OLCF). Called Frontier, it's expected to be the fastest open science supercomputer in the world when it arrives in 2021, and one of the first to offer exascale computing power of 1 exaFLOPS or more. It will also have both AMD CPUs and AMD GPUs at its heart.

ORNL is preparing eight key scientific applications for Frontier, and one of them is CHOLLA, which investigates astrophysics and galaxy formations, one of the first and one of the chief architects of CHOLLA. "But a revolution in astronomy over the last 40 years has been our ability to use numerical simulations to try to understand how the universe is evolving. Unlike most physical sciences where you can conduct experiments, and the time scales of the experiments happen in relevant human lifetimes, in astronomy, things change on much longer timescales. The only way we can get a moving picture of things is by conducting numerical simulations."

"Having access to this exascale machine is a game changer for the kinds of problems that we can simulate."

Evan Schneider, Assistant Professor of Physics and Astronomy at the University of Pittsburgh

CHOLLA was created to provide this time-based analysis, particularly focusing on galaxy formation, with its processing accelerated by GPUs. "The universe is mostly composed of gas and dark matter," says Schneider. "So the two most relevant pieces of physics are fluid



ORNL expands possibilities for plasma physics with open and portable AMD ROCm

Increased fidelity and longer simulations using GPU-accelerated software from AMD

AMD EPYC | INSTINCT

CUSTOMER

Oak Ridge National Laboratory

INDUSTRY

Scientific research

CHALLENGES

Being able to run larger, longer simulations with PIConGPU plasma physics simulations when Frontier supercomputer comes online

SOLUTION

Deploy AMD Instinct™ MI100 GPUs with ROCm™ open software to prepare for Frontier supercomputer

RESULTS

Easy porting to AMD Instinct GPUs with

Oak Ridge National Laboratory (ORNL) has been home to some of the most powerful computational resources used in science for decades. It currently hosts the second most powerful supercomputer in the world, Summit. ORNL has commission a new supercomputer. Powered by AMD CPUs and GPUs, Frontier is expected to be one of the fastest systems in the world when it arrives in 2021. ORNL is readying several key scientific applications that can take advantage of the exascale power the new system will deliver via its Center for Accelerated Application Readiness (CAAR).

Amongst these applications is Particle-In-Cell GPU (PIConGPU). "We are one of the eight teams across the States chosen for the CAAR project to prepare the software stack for Frontier," says Sunita

"Frontier's AMD Instinct GPU computing power will enable us to find answers to questions not accessible before."

Sunita Chandrasekaran, Assistant Professor of Computer and Information Sciences at the University of Delaware

Increasing the kinetic energy of particle acceleration

GPUs have been essential for the computational capability required by PIConGPU, with every advance in computing performance and memory throughput providing a step forward. "For more realistic science cases, higher fidelity models, and more density simulations, we need more compute power per GPU," says Ronnie Chatterjee, Oak Ridge Leadership Computing Facility (OLCF) liaison for PIConGPU. This is because the research team wants to support particle accelerators capable of producing a greater level of kinetic energy than current instruments.

"We want to get to the next generation of compact accelerators," says Dr Alexander Debus, Institute of Radiation Physics, Researcher at Helmholtz

Goethe University Frankfurt delivers broad range of scientific research with AMD

2nd Gen AMD EPYC™ CPUs and AMD Radeon Instinct™ MI50 GPUs power discoveries in particle physics, climate research, digital medicine and more.

AMD EPYC | INSTINCT

CUSTOMER

FIAS Frankfurt Institute for Advanced Studies

INDUSTRY

Theoretical science research

CHALLENGES

Delivering cost-effective HPC modeling and analysis capability to nearly 50 different research groups

SOLUTION

Deploy Servers with AMD EPYC™ CPU and Radeon Instinct™ MI50 GPU

RESULTS

Achieved cost-performance objectives to grow capacity to support more research

Theoretical science pushes boundaries, furthering our understanding of fundamental natural phenomena. The computing platforms used to model and analyze such complexity in detail must be capable of handling everything imaginable, and more. For the Center for Scientific Computing (CSC) at Frankfurt Institute for Advanced Studies (FIAS), located at Goethe University, that demand is amplified by the challenge of supporting high-performance computing needs across nearly 50 independent research groups exploring life sciences, theoretical physics, neuroscience, computer science and systemic risk.

Since the progress of such critical work is constrained by the available compute capabilities, CSC is hard at

"More than 98 percent of processing for the 600 gigabytes per second data stream runs entirely on servers with eight AMD MI50 GPUs that deliver 90% of the total compute performance we need using just the GPUs."

Professor Volker Lindenstruth, chairman of the Board of Directors, FIAS

2nd Gen AMD EPYC™ processors and AMD Radeon Instinct™ MI50 GPU computing accelerators to obtain the combination of performance, advanced I/O capabilities, and operating efficiency needed to accommodate their clients' many varied applications.

And the diversity of research is impressive. There are groups conducting ab initio calculations called lattice quantum chromodynamics (QCD) that model nuclear reactions without free parameters. Another group uses the ultra-relativistic molecular dynamics (UrQMD) simulation code, developed at Goethe University, for applications across particle physics, high energy experimental physics and engineering, shielding, detector design, cosmic ray studies, and medical physics. There are yet other groups exploring quantum chemistry.

"We got most of the porting [of CHOLLA] to HIP to run on AMD hardware done in a few hours."
Reuben Budiardja
Computational Scientist at ORNL

LINK

"We are very appreciative of the fact that AMD has the ROCm open source software stack with the HIP programing model."
Sunita Chandrasekaran, Assistant Professor of Computer and Information Sciences at the University of Delaware

LINK

"We determined that a server built with AMD EPYC CPUs and eight AMD MI50 GPUs delivered ideal cost performance. And, as our GPU code is better optimized, we achieve even greater efficiency."
Professor Volker Lindenstruth
chairman of the Board of Directors,FIAS

LINK

AMD

AMD ROCm™ BROADENS PORTABILITY

PERFORMANCE & PORTABILITY AT HEART OF ROCM

OpenMP® 5.0

OpenCL™

HIP

Open Source LLVM-based Compiler

Single Compiler for All Models

Runs on Variety of Processors, GPUs

Open Source HIP CPU Runtime*

Same HIP Kernel and Source Code Runs on CPU or GPU

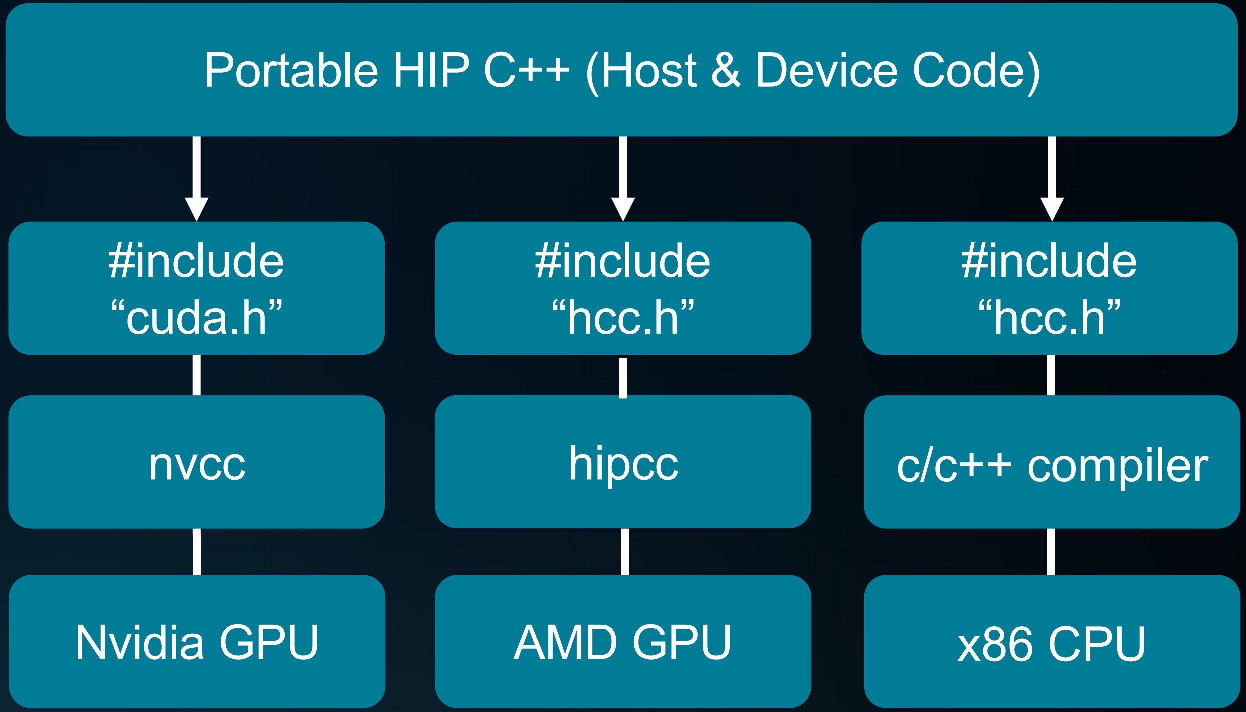
OpenMP on CPU, OpenMP Offload on GPU

*Available separately on github

HIP: HIGH PERFORMANCE AND PORTABLE

C++ runtime API and kernel language that allows developers to create portable applications that can run on AMD's accelerators, x86 CPUs as well as CUDA® devices.

- Syntactically like CUDA
- Most CUDA API calls can be converted in place
- Supports a strong subset of CUDA runtime functionality



HIP – A COMMON KERNEL LANGUAGE FOR DEVICES

CUDA VECTOR ADD FOR GPU

```
__global__ void add(int n, double *x, double *y)
{
  int index = blockIdx.x * blockDim.x + threadIdx.x;
  int stride = blockDim.x * gridDim.x;
  for (int i = index; i < n; i += stride)
    {
    y[i] = x[i] + y[i];
    }
}
```

HIP VECTOR ADD FOR GPU OR CPU

```
__global__ void add(int n, double *x, double *y)
{
  int index = blockIdx.x * blockDim.x + threadIdx.x;
  int stride = blockDim.x * gridDim.x;
  for (int i = index; i < n; i += stride)
    {
    y[i] = x[i] + y[i];
    }
}
```

KERNELS ARE SYNTACTICALLY IDENTICAL

AMD

HIPIFY TOOLS

CONVERTING CUDA® CODE FOR PORTABILITY

Hipify-perl

- Easy to use –point at a directory and it will attempt to hipify CUDA code
- Very simple string replacement technique: may make incorrect translations
- sed -e 's/cuda/hip/g', (e.g., cudaMemcpy becomes hipMemcpy)
- Recommended for quick scans of projects

Hipify-clang

- Requires clang compiler to parse files and perform semantic translation
- More robust translation of the code
- Generates warnings and assistance for additional analysis
- High quality translation, particularly for cases where the user is familiar with the make system

AMD

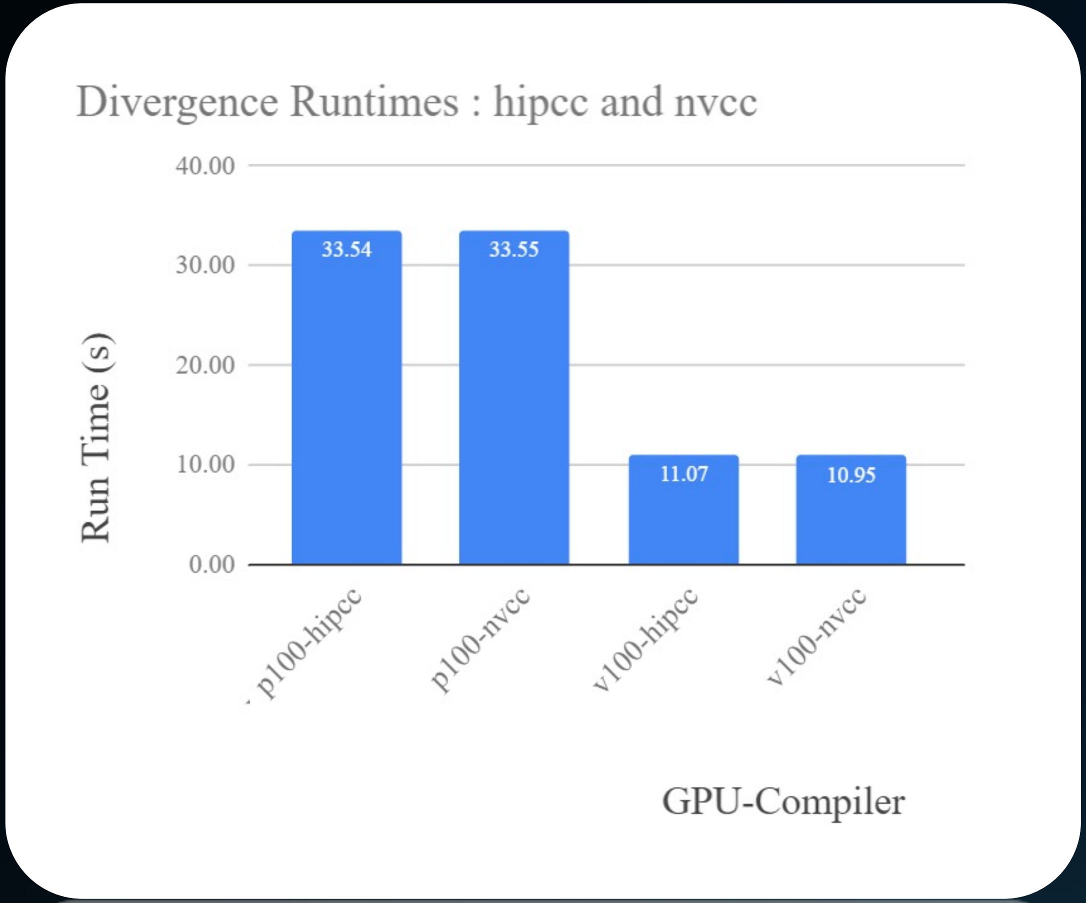
SEAMLESS PORTING FROM CUDA® APIs

CUDA	*HIP*
cudaMemcpyAsync(d_npos,h_npos,sizeof(float4)*SIZE,cudaMemcpyHostToDevice,stream);	hipMemcpyAsync(d_npos,h_npos,sizeof(float4)*SIZE,hipMemcpyHostToDevice,stream);
cudaMemcpyAsync(d_mask,h_mask,sizeof(MASK_T)*cnt,cudaMemcpyHostToDevice,stream);	hipMemcpyAsync(d_mask,h_mask,sizeof(MASK_T)*cnt,hipMemcpyHostToDevice,stream);
calcHHCullenDehnen<<<blocksPerGrid, threadsPerBlock, 0, stream>>>(cnt, SIZE, d_npos, d_mask, rsm);	hipLaunchKernelGGL((calcHHCullenDehnen), dim3(blocksPerGrid), dim3(threadsPerBlock), 0, stream, cnt, SIZE, d_npos, d_mask, rsm);
cudaMemcpyAsync(h_pos,d_npos+(SIZE-cnt),sizeof(float4)*cnt,cudaMemcpyDeviceToHost,stream);	hipMemcpyAsync(h_pos,d_npos+(SIZE-cnt),sizeof(float4)*cnt,hipMemcpyDeviceToHost,stream);
cudaMemcpyAsync(h_mask,d_mask,sizeof(MASK_T)*cnt,cudaMemcpyDeviceToHost,stream);	hipMemcpyAsync(h_mask,d_mask,sizeof(MASK_T)*cnt,hipMemcpyDeviceToHost,stream);

AMD

REAL-WORLD APPLICATION SHOWS POWER OF HIP

HIP CODE DELIVERS SIMILAR PERFORMANCE ON SELF-FLUIDS KERNEL AS CUDA® CODE, ON NVIDIA GPUS



"On the Nvidia systems, the performance of the HIP and CUDA kernels are nearly identical, indicating there are no performance losses from the 'hipification' process."

HIP Performance Comparisons: AMD and Nvidia GPUs
https://journal.fluidnumerics.com/hip-performance-comparisons-amd-and-nvidia-gpus

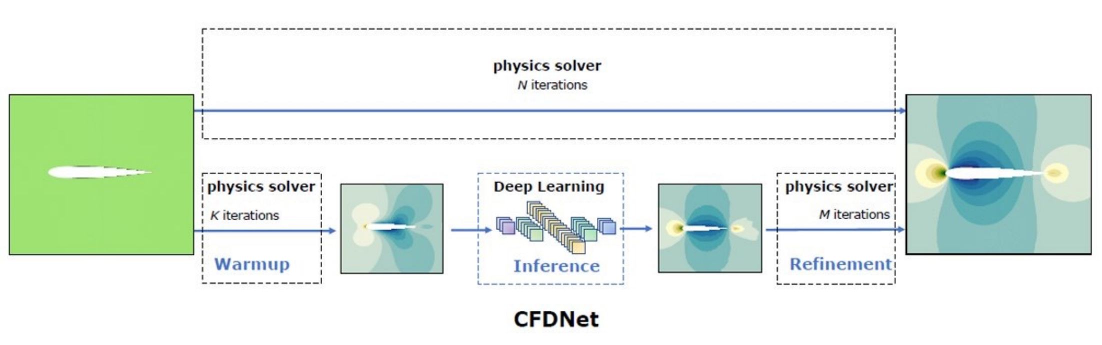
CUDA® COMPARABLE LIBRARIES

CUDA Library	ROCm Library	Comment
cuBLAS	rocBLAS	Basic Linear Algebra Subroutines
cuFFT	rocFFT	Fast Fourier Transfer Library
cuSPARSE	rocSPARSE	Sparse BLAS + SPMV
cuSolver	rocSolver	Lapack Library
AMG-X	rocALUTION	Sparse iterative solvers & preconditioners with Geometric & Algebraic MultiGrid
Thrust	rocThrust	C++ parallel algorithms library
CUB	rocPRIM	Low Level Optimized Parallel Primitives
cuDNN	MIOpen	Deep learning Solver Library
cuRAND	rocRAND	Random Number Generator Library
EIGEN	EIGEN	C++ template library for linear algebra: matrices, vectors, numerical solvers
NCCL	RCCL	Communications Primitives Library based on the MPI equivalents

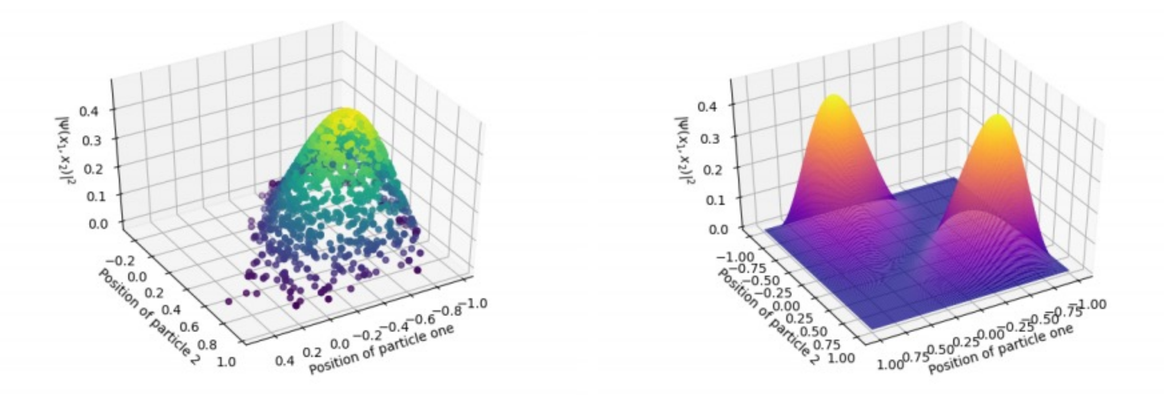
THE CONVERGENCE OF ML/AI + HPC

ENABLING ENHANCED SCIENTIFIC DISCOVERY

- **Neural Network for Surrogate Models: CFDNet**
  - [2005.04485] CFDNet: a deep learning-based accelerator for fluid simulations (arxiv.org), ICS
  - **1.9x-7.4x speedup** without relaxing the convergence constraints of the physics solver
  - Exploring further generalizations to: LES, MD,AMG



- **Improved Variational Monte Carlo**
  - **"Deep Learning on Supercomputers Workshop" at SC20,** https://ieeexplore.ieee.org/abstract/document/9297114
  - 5x speed-ups in select quantum mechanical systems



- COVID19 HPC task force
  - Support for Corona Cluster at LLNL

AMD

PyTorch 1.9

AMD ROCm™ SUPPORT THROUGH BINARIES FROM PyTorch.ORG

INSTALL PYTORCH

Select your preferences and run the install command. Stable represents the most currently tested and supported version of PyTorch. This should be suitable for many users. Preview is available if you want the latest, not fully tested and supported, 1.10 builds that are generated nightly. Please ensure that you have **met the prerequisites below (e.g., numpy)**, depending on your package manager. Anaconda is our recommended package manager since it installs all dependencies. You can also install previous versions of PyTorch. Note that LibTorch is only available for C++.

Additional support or warranty for some PyTorch Stable and LTS binaries are available through the PyTorch Enterprise Support Program.

PyTorch Build	Stable (1.9.0)	Preview (Nightly)	LTS (1.8.1)	
Your OS	Linux	Mac	Windows	
Package	Conda	Pip	LibTorch	Source
Language	Python	C++ / Java		
Compute Platform	CUDA 10.2	CUDA 11.1	ROCm 4.2 (beta)	CPU

Run this Command:

```
pip3 install torch -f https://download.pytorch.org/whl/rocm4.2/torch_stable.html
pip3 install ninja && pip3 install 'git+https://github.com/pytorch/vision.git@v0.10.0'
```

DEEP LEARNING FRAMEWORKS: GET MOST RECENT VERSION TODAY

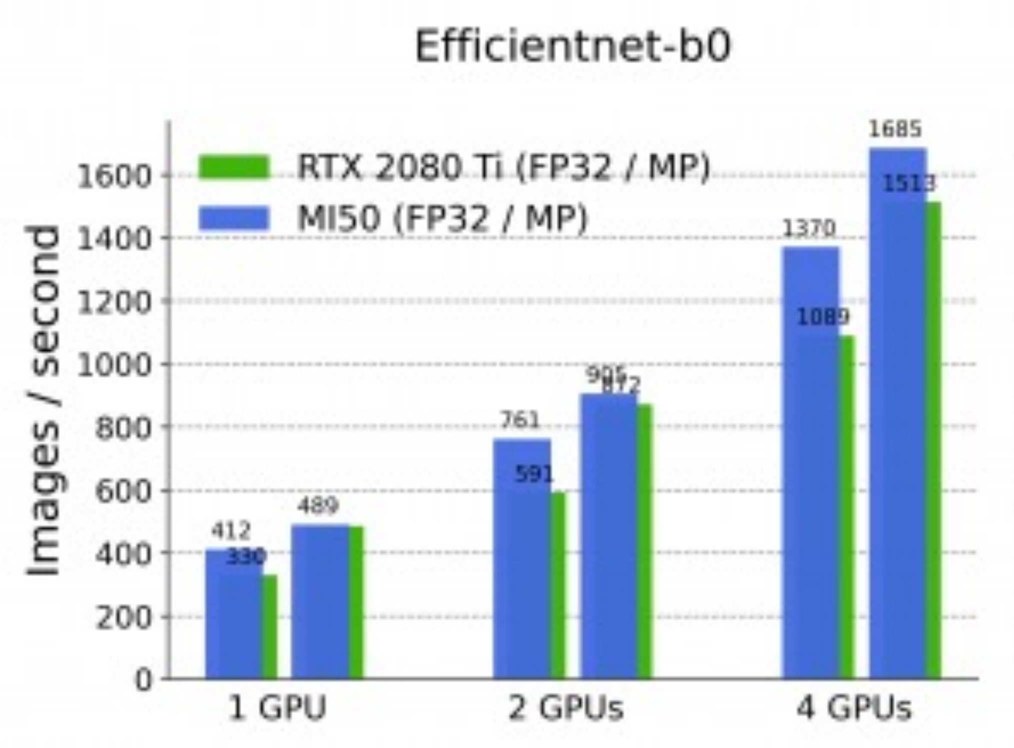
	TensorFlow	PyTorch
Source	https://github.com/tensorflow/tensorflow	https://github.com/pytorch/pytorch
Python PIP Wheel	https://pypi.org/project/tensorflow-rocm/	https://pytorch.org
Docker Container	https://hub.docker.com/r/rocm/tensorflow	https://hub.docker.com/r/rocm/pytorch

AMD

AMD ML GAINING ADOPTION IN ECOSYSTEM

"…we have shown that for some neural network architectures the MI50 is the faster option. The availability of PyTorch with a ROCm backend is a potential game changer for the GPU-for-ML market."

Joris Mollinga
SURF, High Performance Machine Learning Consultant



https://communities.surf.nl/artikel/performance-comparison-of-image-classification-models-on-amdnvidia-with-pytorch-18

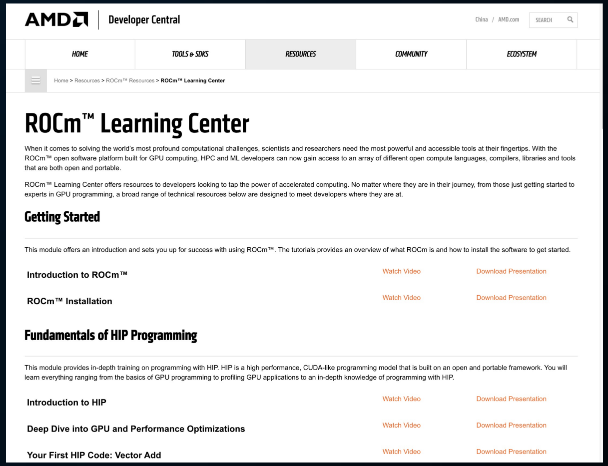
We would like to thank all the folks from **SURF** for providing feedback on our plots. Experiments on the NVIDIA platform were conducted on the **Lisa cluster computer**. Experiments on AMD hardware were conducted on the **Kleurplaat** system within **SURF Open Innovation Lab**.

GETTING STARTED WITH ROCm™ OPEN SOFTWARE PLATFORM

AMD ROCm

ROCm™ Learning Center

Curated videos, webinars, labs and tutorials for developers to learn how to use ROCm
https://developer.amd.com/resources/rocm-resources/rocm-learning-center/



AMD Accelerator Cloud

Private cloud environment for customers and partners to test code and applications on the latest AMD GPUs
Registration Link

AMD INSTINCT™ GPUS & ROCm™

USEFUL WEB RESOURCES

- AMD Instinct GPUs:
  - AMD Instinct™ MI100 GPU page: https://www.amd.com/en/products/server-accelerators/instinct-mi100
  - AMD Instinct™ MI Series Product Page: www.AMD.com/Instinct
  - AMD Instinct™ HPC Solutions Page: https://www.amd.com/en/graphics/servers-radeon-instinct-mi-powered-servers
  - AMD Instinct™ Machine Learning Solutions Page:
  - AMD CDNA Architecture: https://www.amd.com/en/technologies/cdna
  - CDNA WP: https://www.amd.com/system/files/documents/amd-cdna-whitepaper.pdf
  - AMD Infinity Architecture page: https://www.amd.com/en/technologies/infinity-architecture
- AMD ROCm™ open software platform:
  - AMD ROCm™ pages: https://www.amd.com/en/graphics/servers-solutions-rocm
  - ROCm Learning Center https://developer.amd.com/resources/rocm-resources/rocm-learning-center/
  - ROCm DOCs page: https://rocmdocs.amd.com/en/latest/
- HPC & AMD page: www.AMD.com/HPC

For AMD Instinct™ GPU and ROCm™ marketing assets, contact: Guy Ludden Guy.Ludden@AMD.com

Thank you for attending!

Questions?

End Notes

CDNA-04

Calculations by AMD Performance Labs as of Oct 5th, 2020 for the AMD Instinct™ MI100 accelerator designed with AMD CDNA 7nm FinFET process technology at 1,200 MHz peak memory clock resulted in 1.2288 TFLOPS peak theoretical memory bandwidth performance. The results calculated for Radeon Instinct™ MI50 GPU designed with "Vega" 7nm FinFET process technology with 1,000 MHz peak memory clock resulted in 1.024 TFLOPS peak theoretical memory bandwidth performance. CDNA-04

MI100-03

Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak double precision (FP64), 46.1 TFLOPS peak single precision matrix (FP32), 23.1 TFLOPS peak single precision (FP32), 184.6 TFLOPS peak half precision (FP16) peak theoretical, floating-point performance. Published results on the NVidia Ampere A100 (40GB) GPU accelerator resulted in 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16) theoretical, floating-point performance. Server manufacturers may vary configuration offerings yielding different results. MI100-03

MI100-04

Calculations performed by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 accelerator at 1,502 MHz peak boost engine clock resulted in 184.57 TFLOPS peak theoretical half precision (FP16) and 46.14 TFLOPS peak theoretical single precision (FP32 Matrix) floating-point performance. The results calculated for Radeon Instinct™ MI50 GPU at 1,725 MHz peak engine clock resulted in 26.5 TFLOPS peak theoretical half precision (FP16) and 13.25 TFLOPS peak theoretical single precision (FP32 Matrix) floating-point performance. Server manufacturers may vary configuration offerings yielding different results. MI100-04

MI100-05

Calculations performed by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 accelerator at 1,502 MHz peak boost engine clock resulted in 11.535 TFLOPS peak theoretical double precision (FP64) floating-point performance. The results calculated for Radeon Instinct™ MI50 GPU at 1,725 MHz peak engine clock resulted in 6.62 TFLOPS FP64. Server manufacturers may vary configuration offerings yielding different results. MI100-05

MI100-07

Radeon Instinct™ MI50 "Vega 7nm" technology-based accelerators support PCIe® Gen 4.0 providing up to 64 GB/s peak theoretical transport data bandwidth from CPU to GPU per card. Radeon Instinct™ MI50 "Vega 7nm" technology-based accelerators include dual Infinity Fabric™ Links providing up to 184 GB/s peak theoretical GPU to GPU or Peer-to-Peer (P2P) transport rate bandwidth performance per GPU card. Combined with PCIe Gen 4 support providing an aggregate GPU card I/O peak bandwidth of up to 248 GB/s.
MI50 based four GPU hives provide up to 368 GB/s peak theoretical P2P performance. Dual 4 GPU hives in a server provide up to 736 GB/s total peak theoretical direct P2P performance per server.

MI100-08

92.28 TFLOPS peak theoretical bFloat16 precision (BF16) performance based on calculations conducted performed by AMD Performance Labs as of Oct 05, 2020 for the AMD Instinct™ MI100 accelerator at peak 1,502 MHz boost engine clock. Server manufacturers may vary configuration offerings yielding different results. MI100-08

MI100-17

Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak double precision (FP64) theoretical floating-point performance. Nvidia specifications from datasheets at www.nvidia.com/content/en-us/data-center and other sources. MI100-17

AMD

Disclaimer and Attributions

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

AMD